Students' Admission Prediction using GRBST with Distributed Data Mining

Dineshkumar B Vaghela
PhD Scholar
Gujarat Technological University, Chandkheda

Priyanka Sharma, Ph.D
PhD Guide
Gujarat Technological University, Chandkheda

ABSTRACT

Data is the most important asset of any organization which is further processed to produce useful information. Data mining techniques are widely used for industrial sectors to generate the useful pattern helpful for earning more profits and expand business. Since last few years, lots of research works have been done by applying data mining techniques on educational data for improvement in Education System. Data Mining can be useful for predicting such as the students' admission, faculty performance, student performance, identifying the group of students of similar behavior. Very large educational institute's data are geographically spread and increase every year. It is very time consuming and tedious task of processing these large volume of data. In this paper, the new algorithm has been presented with Binary Search Tree which stores the global rules by consolidating the local rules generated at each site. This Global Rule Binary Search Tree (GRBST) can then be used in prediction of Students' admission to college.

Keywords

Binary Search Tree, admission, prediction, distributed data mining, Education System (ES), Big Data, GRBST

1. INTRODUCTION

1.1 Data Mining

"A secret of success is to know something that nobody else knows" said by Aristotle Onassis. Data mining is the process of extracting useful information from given data set, i.e. data mining is the exploration and analysis of large quantities of data in order to discover valid, novel, potentially useful, and ultimately understandable patterns in data. It is the five step process such as Data integration and selection, Preprocessing, Model/pattern construction, Interpretation and knowledge acquisition. DM provides classification, clustering, frequent pattern mining etc techniques.

1.2 Distributed Data Mining

Distributed computing plays an important role in the Data Mining process for several reasons. First, Data Mining often requires huge amounts of resources in storage space and computation time. To make systems scalable, it is important to develop mechanisms that distribute the work load among several sites in a flexible way. Second, data is often inherently distributed into several databases, making a centralized processing of this data very inefficient and prone to security risks. Distributed Data Mining explores techniques of how to apply Data Mining in a non-centralized way. Distributed data mining is a child branch of the data mining. In DDM databases / data warehouses / data servers are located in distributed environment. This environment can be heterogeneous or homogeneous. Analysis and processing of data in distributed environment is a concerned area of research to provide better and faster processing.

1.3 Data Mining Applications

There are many areas where data mining is being used. These areas are listed below.

1. Retail Industries

Retailers collect terabytes upon terabytes of information every day - anything from transactional data, to demographics, to product sales based on seasons. But what do they do with it all once it is neatly organized into a database? The concept of data mining is just as it sounds. Companies drill holes through 0s and 1s to come up with relationships and patterns in customer habits. To a retailer this information can be more valuable than mining for gold, because the results are almost a guarantee.

2. Telecommunication Industries [12]

The telecommunications industry was one of the first to adopt data mining technology. This is most likely because telecommunication companies routinely generate and store enormous amounts of high-quality data, have a very large customer base, and operate in a rapidly changing and highly competitive environment. Telecommunication companies utilize data mining to improve their marketing efforts, identify fraud, and better manage their telecommunication networks. However, these companies also face a number of data mining challenges due to the enormous size of their data sets, the sequential and temporal aspects of their data, and the need to predict very rare events—such as customer fraud and network failures—in real-time.

3. Financial Data Analysis

Financial data analysis is used in many financial institutes for accurate analysis of consumer data to find defaulter and valid customer. For this different data mining techniques can be used. The information thus obtained can be used for Decision making

4. Biological Data Analysis

Biologists are stepping up their efforts in understanding the biological processes that underlie disease pathways in the clinical contexts. This has resulted in a flood of biological and clinical data from genomic and protein sequences, DNA microarrays, protein interactions, biomedical images, to disease pathways and electronic health records. To exploit these data for discovering new knowledge that can be translated into clinical applications, there are fundamental data analysis difficulties that have to be overcome. Practical issues such as handling noisy and incomplete data, processing compute-intensive tasks, and integrating various data sources, are new challenges faced by biologists in the post-genome era.

5. Intrusion Detection

Intrusion detection has become a critical component of network administration due to the vast number of attacks persistently threaten our computers. Traditional intrusion detection systems are limited and do not provide a complete solution for the problem. They search for potential malicious activities on network traffics; they sometimes succeed to find true security attacks and anomalies. However, in many cases, they fail to detect malicious behaviours (false negative) or they fire alarms when nothing wrong in the network (false positive). In addition, they require exhaustive manual processing and human expert interference. Applying Data Mining (DM) techniques on network traffic data is a promising solution that helps develop better intrusion detection systems

6. Other scientific applications

1.4 Educational Data Mining

Educational Data Mining is an emerging discipline of Data Mining which is concerned and focused on developing methods for exploring the unique types of data that come from educational institutes/systems, and using those methodologies for better understanding of students, and the logics which they learn in. As the education is an essential part of the society, more and more colleges and universities are coming in existence and produces huge amount of educational data. Analysis of these data can help education system to improve their system, quality and satisfaction level.

2. BACK GROUND AND LITERATURE REVIEW

Malaya Dutta Borah, Rajni Jindal and Daya Gupta [7] had introduced a heuristic function of C4.5 algorithm and used it to predict the branch of admission for students.

Researcher have also build hybrid recommender system which can be used for the admission process for students[8].

Malaya Dutta Borah, Daya Gupta and Gokul Pandey[9] introduced two modification on C4.5 algorithm and compared its results with original C4.5 algorithm. They have done branch prediction and future grade prediction in their work.

A new Binary Search Tree algorithm is introduced by Sachin Makwana, Dinesh Vaghela and Priyanka Sharma[11] which is concerned with creating a BST whose nodes store the attribute values based on rule generation at each site. The prediction will be faster with this form of BST.

Simon Fong, Yain-Whar Si, Robert P. Biuk Aghai [10] had used back-propogation algorithm and C4.5 algorithm for the student admission process.

Educational data mining is the most recent research area which extracts useful and unknown patterns from data warehouse for better reorganization and increased performance of the student learning process [1]. Miren Tanna [2] gave idea about Decision Support System which is based on CET (common eligibility test) and introduce a new idea about offset value.

Suresh kumar Yadav and Saurabh Pal [3] have used decision tree method to predict the student performance based on which they can calculate the eligibility of the student for the course of MCA. They have used entropy, GINI Index and classification error to measure degree of impurity of a given data set. Information GAIN is used to determine best attribute for a particular joint/point in the tree. Separation/division criteria are formatted using this value.

Neeraj Bhargava, Anil Rajput and Pooja Shrivastava [4] used different algorithms like CHAID ,CART, J48, C4.5 and C5.0/ They have formed an excel sheet including following attributes.

Course, Branch, Gender, Category, Class, Income,

Date of Admission, Minority/Non-Minority

They applied different decision tree algorithms on this excel sheet in data mining tool Weka. They choose the algorithm which gives the best result for the given data set.

3. PROBLEM IDENTIFICATION

Existing systems are for local environment and that may be too good in local environment. But what if, Data are stored at different servers? What if System is being developed for a university which has many affiliated colleges?

There are many collages under one university. so it is difficult to maintain the single database for all the collages. This can be referred as Big Data as its size increases tremendously with time. Distributed data mining is the solution for Big Data problem.

Doing same work in Distributed Environment will give result slower with compare to local environment. Accuracy may also be reduced for the same algorithm if datasets/databases are scattered over different web servers. Providing a system in distributed environment with same or higher accuracy rate is the solution to the problem. Network security and other overhead is not in concern right now. The main focus is to reduce the processing time of the dataset in distributed environment.

4. RESEARCH DESIGN AND METHODOLOGY

4.1 Proposed Framework

In the proposed framework of the system there will be two sites called Local Site and Remote Site. Local Site consists of Application layer, Middleware and Meta Data directory. Remote Site consist Data warehouses at different servers.

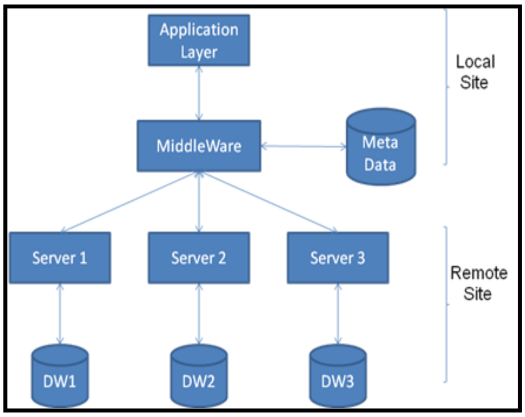


Fig. 1. Proposed Framework for Distributed Environment

4.2 Working of System

As a general scenario, Admission in colleges / universities takes place at every one year. Because of this reason there will be no change bigger change in academic data in between an academic year. So, training phase will done once a year, whereas testing phase could happen as per requirements.

(1) Training Phase: Training of the available data will done either by admin / automatically once a year as there will be no major changes in the data and thus it will not affect the decision making.

In training phase, J48 algorithm will be applied to all the local sites. Rules generated at a local site ,i.e. at local server, will fetched to middleware and consolidation will take place to form a global rule. Once Global rules are generated, it will be given to all the servers for future testing phase.

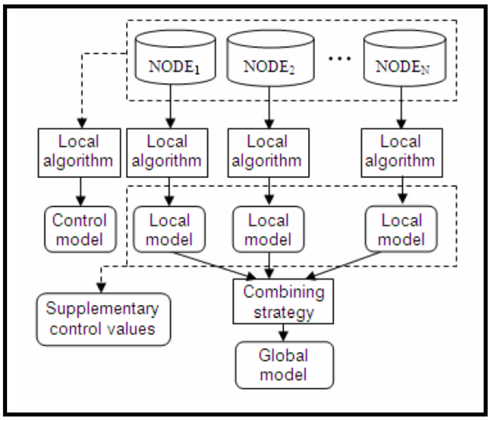


Fig. 2. Distributed Data Mining Scheme[5]

(2) Testing phase: Testing phase can be done by users and admin as well. Users can interact with the system with the help of application layer which is User Interaction system of our application. They can enter data for single user or multiple data in the provided format at our site. The format will be .xls, .xlsx or .csv.

Once user has entered his query, it will come to metadata directory via middleware. If query is fresh i.e. it is for the first time then it will forwarded to all server. Whatever result will come from the servers are going to be same as all servers have global rules, so this result will be displayed to the user.

In the case of multiple queries at a time, the uploaded file will be subdivided horizontally into n parts, where n is the number of servers in our system, and each subset will sent to different server.

After processing this file, result will come to middleware. Middleware will wait for results from all servers. After getting all results from all servers, middleware will merge the results to form a single file and give it to the user as output file.

4.3 Consolidation Technique

Consolidation of rules at middleware is a difficult task. In this paper the Conflicting Rules separations and Sub Rules elimination strategies [5] have been used. In the conflicting rules separations strategy, attributes of the rules which are conflicting each other are compared and the most effective rule is considered while others are deleted. Whereas in the sub rules elimination strategy, rule which is a subset of another rule is eliminated. This can be done by two ways.

(1) using If..Then.. rules and

(2) using Decision Table.

(1) Using If.. Then... rules format.

Example No.1

RULE 1: if MERIT_MARKS is (>129) and CITY is (VADODARA) and HSC is (84) and CATEGORY is (OPEN) and ACPC_RANK is (<1253) then COLLEGE = BVM

RULE 2: if MERIT_MARKS is (>129) and CITY is (VADODARA) and HSC is (80) and CATEGORY is (OPEN) AND ACPC_RANK is (<1253) then COLLEGE = PARUL

Here, Rule 1 and Rule 2 are considered from different servers. It is not possible that for one merit marks predicted collage will be different. hence at this time consider the other parameter like CATEGORY and HSC percentage. Here CATEGORY is same so check for HSC percentage. If Merit_Marks are same and CATEGORY is different than keep the both rules. Now according to HSC percentage higher the percentage more the probability, so compare the HSC percentage of rules. Here, Rule 1's HSC percentage is greater than the Rule 2's. so we'll delete the Rule 2 and consider Rule 1 for prediction.

Example No. 2

RULE 1: if MERIT_MARKS is (>120) and CITY is (VADODARA, ANAND, SURAT, AHMEDABAD) and HSC is (84) and CATEGORY is (OPEN) and ACPC_RANK is (<1253) then COLLEGE = PARUL

RULE 2: if MERIT_MARKS is (>120) and CITY is (ANAND,VADODARA) and HSC is (84) and CATEGORY is (OPEN) and ACPC_RANK is (<1253) then COLLEGE = PARUL

Here, Rule 1 and Rule 2 have similar attributes values and similar conclusion. So, Rule 2 is subset of Rule 1. Now delete Rule 2 as it is subset of Rule 1 so it is not necessary for prediction.

(2) Using Decision Table

Decision Table is an important format which helps in classification and prediction process. Decision Table can be generated from decision tree by taking each traversing path as a raw of the table. After converting decision tree into decision table same methodology can be applied to two decision table i.e. Conflicting Rules Separation and Sub Rule Elimination techniques.

Table 1. Decision Table generated from decision tree.

Rank	College	Marks	CITY	HSC	Cast	ACPC Rank
1	BVM	>129	Vadodara	84	OPEN	<1253
2	PARUL	>129	Vadodara	80	OPEN	<1253

Decision Table like above will be created at every server which will then be fetched to the middleware. At middleware comparison of two decision table will be done and based of conflicting rule separation and sub rule elimination methodology rows will be deleted from the decision table and then it will be merged to form a global decision table. Using this global decision table Binary Search Tree can be created which will then be used for decision making.

4.4 Binary Search Tree Construction

Binary Search Tree construction is important stage in our proposed work. Binary search tree can be created by taking the MERIT_MARKS as key value. Steps for creating binary search tree is as follows.

1) *Create Data structure from if..then rules or from rows of decision table which contains all required elements and set Merit_Marks as its primary key element.*

2) Sort all data structure, find its median and select appropriate structure to create root node.

3) now select another structure and check whether it is greater than or less than the root node.

a) set it as left leaf node if it is less than root node.

b) set is as right leaf node if it is greater than root node.

c) set is as either right leaf node or left leaf node if it is equals to root node.

4) Repeat steps 2, 3 for all remaining structures.

For Example,
if...then..rules

Rule 1: if MERIT_MARKS > 195 and CAT = gen and HSC = 84 then COLLEGE = Parul
Rule 2: if MERIT_MARKS <= 195 and MERIT_MARKS = 190 and CAT = gen and HSC = 75 then COLLEGE = SVIT
Rule 3: if MERIT_MARKS >195 and MERIT_MARKS = 200 and CAT = gen and HSC = 86 then COLLEGE = BABARIA
Rule 4: if MERIT_MARKS <= 195 and MERIT_MARKS = 190 and CAT = sc and HSC = 75 then COLLEGE = Parul
Rule 5: if MERIT_MARKS >195 and MERIT_MARKS = 200 and CAT = sc and HSC = 80 then COLLEGE = SVIT
Rule 6: if MERIT_MARKS >195 and MERIT_MARKS = 200 and CAT = gen and HSC = 74 then COLLEGE = SIGMA
Rule 7: if MERIT_MARKS <=195 and MERIT_MARKS = 192 and CAT = sc and HSC = 79 then COLLEGE = BVM

Decision Table

Table 2. Decision Table to generate BST

Sr. No.	College	Merit_Marks	CAT	HSC
1	Parul	>195	gen	84
2	SVIT	=190	gen	75
3	BABARIA	=200	gen	86
4	Parul	=190	sc	75
5	SVIT	=200	sc	80
6	SIGMA	=200	gen	74
7	BVM	=192	sc	79

In order to create binary search tree, first select Rule 1 and create the root node. Now select second rule, which is less than the root node so set it as left leaf node. Third rule, which is greater than root node, will set as right leaf node. This process will continue for all rules.

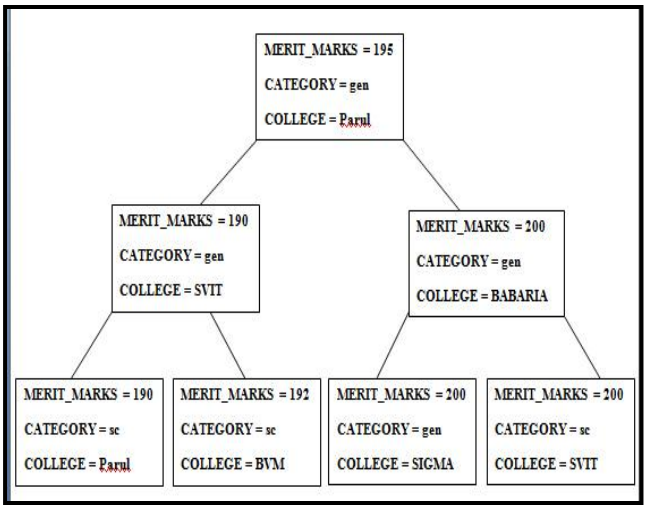


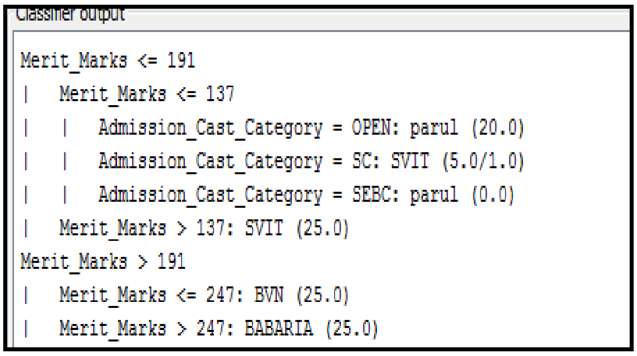
Fig. 3. Generation of Binary Search Tree from rules

5. RESULTS

The time complexity of generating the Binary Search Tree from the Decision table is very less and also this BST has efficient time complexity to predict the result.

5.1 Testing Results on local machine

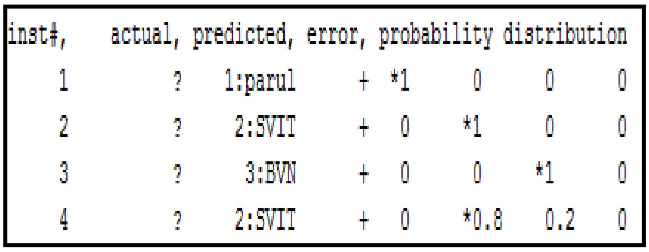
For the testing purpose the dataset of 100 values has been chosen, which contains information of past data of students. This dataset is supplied in WEKA as training set and J48 (which is a java implementation of C4.5) algorithm is applied. Following is the training model which is generated after applying J48 algorithm.



```
Merit_Marks <= 191
|   Merit_Marks <= 137
|   |   Admission_Cast_Category = OPEN: parul (20.0)
|   |   Admission_Cast_Category = SC: SVIT (5.0/1.0)
|   |   Admission_Cast_Category = SEBC: parul (0.0)
|   Merit_Marks > 137: SVIT (25.0)
Merit_Marks > 191
|   Merit_Marks <= 247: BVN (25.0)
|   Merit_Marks > 247: BABARIA (25.0)
```

Fig. 4. Rules(i.e. Decision Tree) generated by J48 algorithm

The 4% of this dataset have supplied as the test dataset in WEKA and following result is generated.



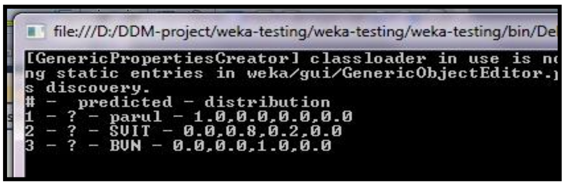
```
inst#,    actual, predicted, error, probability distribution
     1        ?    1:parul      +  *1      0      0      0
     2        ?     2:SVIT      +   0     *1      0      0
     3        ?      3:BVN      +   0      0     *1      0
     4        ?     2:SVIT      +   0   *0.8    0.2      0
```

Fig. 5. Prediction Result generated by J48 algorithm

In above figure, prediction result is displayed for 4 instances. In probability distribution four columns refers to the parul, SVIT, BVN and BABARIA respectively. Actual column displays the actual value of the instance and predicted column displays predicted value for the instance.

Prediction for first 3 instances are parul, SVIT and BVN respectively, and probability of predicted value is 100%. Here, for 4th instance predicted value is SVIT and probability is divided in two columns i.e. SVIT and BVN. As the value of SVIT column is higher than BVN column, final result for the fourth instance is SVIT. Probability of getting admission in SVIT college is 80% while getting admission in BVN college is 20%.

A console system has been implemented in local environment for testing purpose. In this system the same training set has been supplied and used 3% of it as test set. Figure 6 shows the result after supplying the test set. Here each college is separated by a comma under distribution tag. Predicted results for attribute 1,2 and 3 are parul, SVIT and BVN with probability of 100%, 80% and 100% respectively.



file:///D:/DDM-project/weka-testing/weka-testing/weka-testing/bin/De

```
[GenericPropertiesCreator] classloader in use is no
ng static entries in weka/gui/GenericObjectEditor.
s discovery.
# -  predicted - distribution
1 - ? - parul - 1.0,0.0,0.0,0.0
2 - ? - SUIT - 0.0,0.8,0.2,0.0
3 - ? - BVN - 0.0,0.0,1.0,0.0
```

Fig. 6. Test Results in a Console Application

5.2 Comparison of Results

As shown in fig. 7. The student data set of varying size have been processed with the existing algorithm at each local site and our proposed algorithm which generates the Global Rule BST. There has been huge difference of time, the proposed algorithm seems much faster than the existing one. Further this GRBST can also be more efficient for prediction.

Number of Records (ten thousands)	Total Time (Local Site: ms)
1	0.0157
5	0.093
10	1.638
50	9.086
100	20.6871

Total Number of Sites: 3

Number of Records (Ten thousands)	Total Time (ms)
1	0.013
5	0.07
10	0.168
50	0.905
100	2

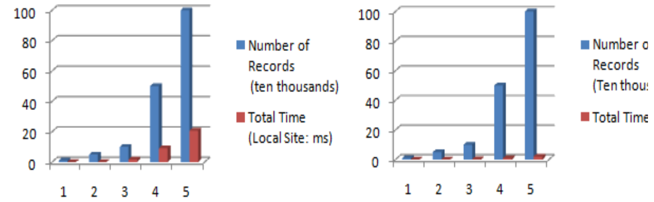


Fig. 7. Time (ms) Required to generate the Model

6. CONCLUSION AND FUTURE WORK

Educational data mining is an emerging and uncovered area for research work of Data mining. Data mining techniques can be useful in deriving patterns from educational data, and this pattern can be useful to improve Education System. This research work can be helpful for any Educational System. Security is the major concern, so in future the security parameters can also be considered.

7. REFERENCES

[1] Priyanka Saini, Ajay Kumar Jain, " Prediction using Classification Technique for the Students' Enrollment Process in Higher Educational Institutions", IJCA, vol 84-no 14, December 2013

[2] Miren Tanna, "Decision Support System for Admission in Engineering Colleges based on Entrance Exam Marks", IJCA, vol 52-no 11, August 2012

[3] Surjeet Kumar Yadav, Saurabh Pal, " Data Mining Application in Enrollment Management: A Case Study", IJCA, vol 41- no 5, March 2012

[4] Neeraj Bhargava, Anil Rajput, Pooja Shrivastava, "Mining higher educational students data to analyze studentÊs admission in various discipline" , BJDMN1, 2010

[5] Marcin Gorawski, Ewa Pluciennik-Psota, "Distributed Data Mining Methodology with Classification model example" , SPRINGER, 2009

[6] Sachin H Makwana, Dinesh B Vaghela, Priyanka Sharma, " A DECISION SUPPORT APPLICATION FOR STUDENT ADMISSION PROCESS BASED ON PREDICTION IN DISTRIBUTED DATA MINING", IC-IKR-EMS, December 2014

[7] Malaya Dutta Borah, Rajni Jindal, Daya Gupta, "Application of Knowledge based decision technique to prediction student enrollment decision", IEEE, 2011, pg. 180 - 184

[8] Abdul Hamid M Raghab, Abdul Fatah S. Mashat, Ahmed M Khedra, "HRSPCA: Hybrid Recommender System for Predictiing College Admission", IEEE, 2012, pg. 107- 113

[9] Daya Gupta, Malaya Dutta Borah, Gokul Pandey, " "EDU-OPT"-A Decision Support Tool", ISTE, 2013

[10] Simon Fong, Yain-Whar Si, Robert P. Biuk Aghai, "Applying a Hybrid Model of Neural Network and Decision Tree Classifier for Predicting University Admission", IEEE, 2009

[11] Sachin H Makwana,Dinesh B Vaghela, Priyanka Sharma, "An Effective Approach for Student Admission Prediction using Binary Search Tree with Distributed Data Mining", ICIIECS, 2015, pg.109-113

[12] Gary M. Weiss Fordham University, USA Data Mining in the Telecommunications Industry- Copyright © 2009, IGI Global